Chronic Kidney Disease Prediction Using Machine Learning

Sathiya Priya S
PG Scholar,
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore.

Suresh Kumar M
Professor,
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore.

Abstract **-** Chronic Kidney Disease prediction is one of the most important issues in healthcare analytics. The most interesting and challenging tasks in day to day life is prediction in medical field. In this paper, we employ some machine learning techniques for predicting the chronic kidney disease using clinical data. We use three machine learning algorithms such as Decision Tree(DT) algorithm, Naive Bayesian (NB) algorithm. The performance of the above models are compared with each other in order to select the best classifier in predicting the chronic kidney disease for given dataset.

Index Terms – Machine Learning; Chronic Kidney Disease; Prediction.

1. INTRODUCTION

Computer vision has been one of the most remarkable breakthroughs for the machine learning and in particular for active healthcare applications. Machine learning allows to build the models to quickly analyze data and deliver results for the given data. Healthcare service providers can make better decisions on patient's disease diagnosis and treatment for the particular disease with the help of machine learning. The massive quantities of data are analysed using machine learning. It delivers faster and more accurate results in order to identify the risks, it may also require additional time and resources to train it proper manner.

Supervised machine learning algorithms can applied to predict the future events with the help of what has been learned in the past to new data using labeled examples. First the known training dataset is analyzed, with that the learning algorithm produces an inferred function to make predictions about the output values.

After sufficient training the system is able to provide targets for any new inputs. Supervised learning algorithms uses patterns to predict label values on additional unlabeled data. As per Fig 1.1 Machine learning algorithms are classified in two types they are supervised machine learning algorithms and unsupervised machine learning algorithms. Supervised machine learning algorithms are based on input-output pairs patterns [1]. These algorithms aims to predict output values based on given input values. Supervised machine learning algorithms mainly focuses on classification and regression.

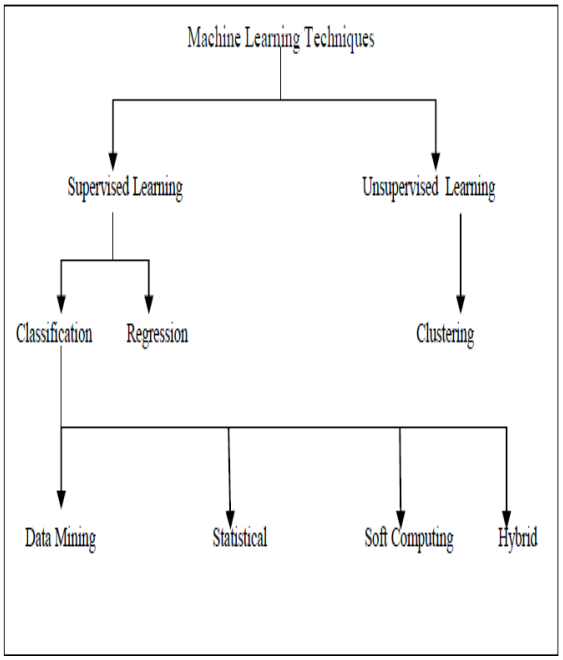

Fig 1.1 classification of machine learning techniques.

2. RELATED WORK

Chronic kidney disease will harm kidneys and reduce its ability to keep our body in a healthy condition. The risk of having heart and blood vessel disease increases due to kidney disease. For the considered dataset the Random Forest has produced better prediction performance in terms of classification accuracy, AUC respectively. The classification performances of the classifier is analyzed with the standard performance parameters, such as: Accuracy, Specificity, Sensitivity, Precision [2].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The machine learning algorithms behaviour were determined on a set of data mining indicators has a relative effect on the models. Knowledge discovery from the wide databases is known as Data mining. Besides studying the existing available Clinic Foundation Heart Disease dataset, 600 clinical records collected from a leading Chennai based diabetes research centre. The application of Data mining technique is a good method for different analysis of medical data. [3]. The chronic disease is predicted with the clinical data using machine learning algorithms. The machine learning algorithms used here includes K-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR). The predictive models are constructed for the taken dataset and the best classifier is predicted using the performance of the models. SVM classifier gives the highest accuracy and has the highest sensitivity after training and testing [4]. The healthcare industry collects large amounts of medical data and for effective decision making the data need to be mined to discover hidden information. Based on the clinical data of patients the heart disease prediction system can assist medical professionals in the prediction of heart disease. The marginal success is achieved with the predictive model for heart disease patients and there is a need for more complex models to increase the accuracy of prediction of the early stages of heart disease [5]. Classification technique normally divides the data into two different data sets one is training set and the other is testing sets. Every occurrence in the training set contains one target variable and several attributes or features. The training data is used to develop a model in SVM, their features which successively predicts the target values of the test data given only the attributes of the input test data. Random Forest is a collection of a group of tree predictors which uses classification technique [16].

3. PROPOSED SYSTEM

The proposed system deals with the prediction of chronic disease from the clinical data. The healthcare generates large data, so it is necessary to collect this data and effectively use it for analysis, prediction, and treatment.

A classification model draws some conclusion from observed values. In classification model one or more inputs are used to predict the value of one or more outcomes. The dataset is applied with the labels which are outcomes. In a supervised machine learning algorithms, the classification algorithm uses the training dataset. classification predicts the categorical class labels whereas the prediction predicts the unknown or missing values.

A decision tree is a tree structure in which internal nodes i.e., non leaf nodes denotes a test on an attribute. Branches denotes the outcomes of tests. Leaf nodes i.e., terminal nodes hold class labels. Root node is the topmost node in the decision tree. A path is traced to leaf node from root node which holds the prediction for the given tuple.

Any domain knowledge or parameter setting is not required for the construction of a decision tree. Decision tree can handle high dimensional data and it can be understood by humans easily. Learning and classification are simple ,fast and it has good accuracy.

Naive Bayes classifier is a powerful algorithm for the classification task. Even with working on a data set with millions of records with some attributes, Naive Bayes approach is best to use. Naive Bayes classifier uses the Bayes Theorem. For each class it predicts membership probabilities such as the probability that given record or data point belongs to a particular class. ‘A’ denotes prior event and ‘B’ denotes dependent event, Bayes’ theorem can be given as

Prob(AgivenB)= Prob(AandB)/Prob(B)

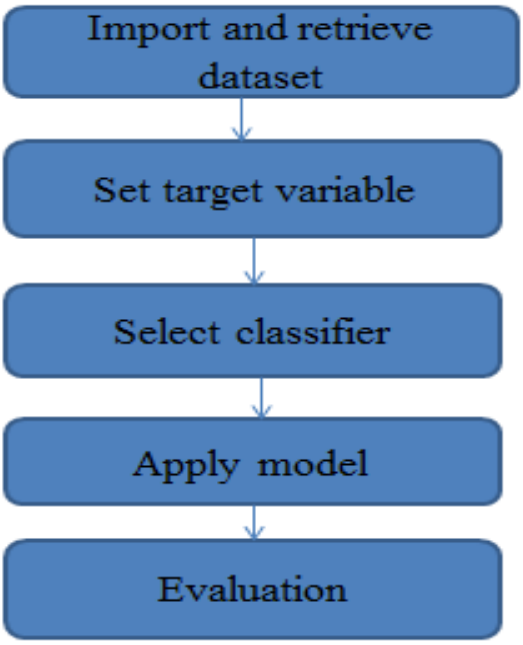

Fig 3.1 The process model to predict chronic kidney disease.

The process steps for Fig 3.1 is as follows: First the dataset is imported and retrieved using the basic steps. Then, the target variable is fixed. Next, the classification method is selected to predict the chronic kidney disease.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Then, the model for the classifier is applied to get the prediction results. Finally, the results of the different classifier are evaluated using some parameters.

4. DATASET

The proposed system uses the dataset taken from the UCI Machine Learning Repository named Chronic Kidney Disease has 25 attributes, 11 numeric and 14 nominal. Total 400 instances of the dataset is used for the training to prediction algorithms, out of which 250 has label chronic kidney disease (CKD) and 150 has label non chronic kidney disease (NOTCKD). The attributes in the dataset are age, bp, sg, al, su, bc, pc, pcc, ba, bgr, bu, sc, sod, pot, hemo, pcr, wc, rc, htn, dm, cad, appet, pe, ane, classification.The dataset is divided into two groups, one for training and another for testing. The ratio of training and testing data is 70% and 30% respectively.

5. RESULT AND DISCUSSION

The machine learning methods described are trained to predict the chronic kidney disease. Two classifier methods are used in this decision tree and naive bayes . The experiments are constructed on R tool. In this work , the performance is measured by sensitivity, specificity and accuracy described as follows.

Accuracy (ACC) is the overall success rate of the classifier defined as

$$ACC = (TP + TN) / (TP + FP + TN + FN)$$

Sensitivity or the true positive rate (TPR) which is defined as the fraction of positive instances predicted correctly by the model defined as

$$Sensitivity = TP / (TP + FN).$$

Specificity is the true negative rate (TNR) which is defined as the fraction of negative instances predicted correctly by the model defined as

$$Specificity = TN / (FP + TN).$$

Where
TP - the number of true positives.
TN - the number of true negatives.
FP - the number of false positives.
FN - the number of false negatives.

With the help of True Positive (TP) and True Negative (TN) the performance of the classifications model is evaluated. The machine learning techniques used are trained and tested separately in this work. The 10-fold cross validation is used to train and test the machine learning models in this work and the average results are shown in table 5.1.

Table 5.1 Performance evaluation for Decision Tree and Naive Bayes classification techniques.

Techniques used	Accuracy	Sensitivity	Specificity
Decision Tree	99.25%	99.20%	99.33%
Naive Bayes	98.75%	98%	98.75%

From table 5.1 by comparing the decision tree method and naive bayes method the accuracy of decision tree method is relatively higher than the naive bayes method. The decision tree method can be adopted since it has the accuracy of 99.25% in prediction of chronic kidney disease.

6. CONCLUSION

The prediction of chronic kidney disease is very important and now-a-days it is the leading cause of death. The performance of Decision tree method was found to be 99.25% accurate compared to naive Bayes method. Classification algorithm on chronic kidney disease dataset the performance was obtained as 99.33% Specificity and 99.20% Sensitivity. We are also further working on enhancing the performance of prediction system accuracy in neural network and deep learning algorithm .

7. REFERENCES

[1] Madhura Rambhajani, Wyomesh Deepanker, Neelam Pathak (2015), "A Survey On Implementation Of Machine Learning Techniques For Dermatology Diseases Classification", International Journal of Advances in Engineering & Technology.

[2] Manish Kumar (2016), "Prediction Of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm", International Journal of Computer Science and Mobile Computing , Vol. 5, Issue. 2, pg.24 – 33.

[3] K. R. Anantha Padmanaban and G. Parthiban (2016), "Applying Machine Learning Techniques For Predicting The Risk Of Chronic Kidney Disease" Indian Journal of Science and Technology, Vol. 9(29).

[4] Anusorn Charleonnan, Thipwan Fufaung, Tippawan Niyomwong, Wandee Chokchueypattanakit, Sathit Suwannawach, Nitat Ninchawee, (2016) "predictive analytics for chronic kidney disease using machine learning techniques", The 2016 Management and Innovation Technology International Conference .

[5] Jaymin Patel, Prof.Tejal Upadhyay, Dr. Samir Patel, (2016) "Heart Disease Prediction Using Machine learning and Data Mining Technique", International Journal Of Computer Science & Communication, Vol. 7, No. 1, pp.129 – 137.

[6] Jerez, J. M.,Molina, I., Garcia-Laencina, P. J., Alba, E., Ribelles, N.,Martin,M., and Franco, L, (2010) "Missing data imputation using statistical and machine learning methods in a real breast cancer problem". Artif. Intell.,Med. 50, 2, 11–11.

[7] A. Asuncion and D. J. Newman. (2007). UCI Machine Learning Repository [Online]. Available: http://www.ics.uci.edu/~mlearn/ MLRepository.html

[8] Witten H, Ian H, (2011) "Data mining: practical machine learning tools and techniques", Morgan Kaufmann Series in Data Management Systems.

[9] P. B. Jensen, L. J. Jensen, and S. Brunak, (2012) "Mining electronic health records: towards better research applications and clinical care," Nature Reviews Genetics, Vol. 13, no. 6, pp. 395–405.

[10] J. C. Ho, C. H. Lee, and J. Ghosh , (2014) "Septic shock prediction for patients with missing data," ACM Transactions on Management Information Systems (TMIS),Vol. 5, no. 1, p. 1.

[11] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, (2015) "A relative similarity based method for interactive patient risk prediction," Data Mining and Knowledge Discovery, vol. 29, no. 4, pp. 1070–1093.

[12] Hudson Fernandes Golino, Liliany Souza de Brito Amaral, Stenio Fernando Pimentel Duarte, Cristiano Mauro Assis Gomes, Telma de Jesus Soares, Luciana Araujo dos Reis, and Joselito Santos, (2014) "Predicting increased blood pressure using machine learning", Journal of obesity.

[13] Tina Patil, R & Sherekar, SS, (2013) "Performance Analysis of Naive bayes and J48 Classification Algorithm for Data Classification", International Journal of Computer Science and Applications, vol. 6, no.2, pp. 256-261.

[14] D. M. F. bin Othman and T. M. S. Yau, (2007) "Comparison of Different Classification Techniques Using WEKA for Breast Cancer," in 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006, F. Ibrahim, N. A. A. Osman, J. Usman, and N. A. Kadri, Eds. Springer Berlin Heidelberg, pp. 520–523.

[15] Taiwo Oladipupo Ayodele, (2010) "Types of Machine Learning Algorithms", New Advances in Machine Learning, Yagang Zhang (Ed.), InTech.

[16] Ashfaq Ahmed K, Sultan Aljahdali, Nisar Hundewale and Ishthaq Ahmed K , "Cancer Disease Prediction With Support Vector Machine And Random Forest Classification Techniques", IEE Cybernetics 2012.